Radiomics in paediatric neuro-oncology

Fetit, Ahmed E.; Novak, Jan; Rodriguez, Daniel ; Auer, Dorothee P.; Clark, Christopher A.; Grundy, Richard G.

Published in:
NMR in Biomedicine

DOI:
10.1002/nbm.3781

Publication date:
2018

Document Version
Peer reviewed version

Title

Radiomics in Paediatric Neuro-Oncology: A Multicentre Study on MRI Texture Analysis.

Author names and affiliations

Dr Ahmed E. Fetit [a, b]
Email: ahmedfetit@gmail.com

Dr Jan Novak [b, c]
Email: j.novak@bham.ac.uk; Telephone: +441213338744

Dr Daniel Rodriguez [d]
Email: daniel.rodriguez@nuh.nhs.uk

Professor Dorothee P. Auer [d, e]
Email: dorothee.auer@nottingham.ac.uk; Telephone: +44115 823 1178

Professor Christopher A. Clark [f]
Email: christopher.clark@ucl.ac.uk; Telephone: +44207 905 2286

Professor Richard G. Grundy [d, e]
Email: richard.grundy@nottingham.ac.uk; Telephone: +44115 823 0620

Professor Andrew C. Peet * [b, c]
Email: a.peet@bham.ac.uk; Telephone: +441213338711

Professor Theodoros N. Arvanitis* [a, b]
Email: t.arvanitis@warwick.ac.uk; Telephone: +442476151601

[a] Institute of Digital Healthcare, WMG, University of Warwick, Coventry, CV4 7AL, UK
[b] Birmingham Children's Hospital NHS Foundation Trust, Birmingham, B4 6NH, UK
[c] Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, B15 2TT, UK
[d] University of Nottingham, Nottingham, NG7 2RD, UK
[e] University Hospital Nottingham, Nottingham, NG7 2UH, UK
[f] Institute of Child Health, University College London, London, WC1N 1EH, UK

*These authors made an equal contribution to the work.

Correspondence

Professor Theodoros N. Arvanitis
Institute of Digital Healthcare, International Digital Laboratory, WMG,
University of Warwick, Coventry, CV4 7AL, UK.
Email: t.arvanitis@warwick.ac.uk
Telephone: +442476151601

Financial support

AEF and TNA would like to thank WMG at University of Warwick for financial support. JN is supported by Help Harry Help Others. ACP is supported by an NIHR Research Professorship. All authors acknowledge the support received from CRUK and EPSRC Cancer Imaging Programme at the Children's Cancer and Leukaemia Group (CCLG) in association with the MRC and Department of Health (England) (C7809/A10342).

Word count

7992

Abstract summary

Motivation: Brain tumours are the most common solid cancers in children in the UK and are the most common cause of cancer deaths in this age group. Despite current advances in magnetic resonance imaging (MRI), non-invasive diagnosis of paediatric brain tumours is yet to find its way in routine clinical practice. *Radiomics*, the high-throughput extraction and analysis of quantitative image features (e.g. texture), offers potential solutions for tumour characterisation and decision support.

Aim and Methods: In the search for diagnostic oncological markers, the primary aim of this work was to study the application of MRI texture analysis (TA) for the classification of paediatric brain tumours. A multi-centre study was carried out, within a supervised classification framework, on clinical MR images and a support vector machine (SVM) was trained with 3D textural attributes obtained from conventional MRI. To determine the cross-centre transferability of TA, assessing how SVM performs on unseen datasets was carried out through rigorous pairwise testing. The study also investigated the nature of features that are most likely to train classifiers that can generalise well with the data. Finally, the issue of class imbalance, which arises due to some tumour types being more common than others, was explored.

Results: For each of the tests carried out through pair-wise testing, optimal area under the ROC curve (AUC) ranged between 76% and 86%, suggesting that the model was able to capture transferable tumour information. Feature selection results suggest that similar aspects of tumour texture are enhanced by MR images obtained at different hospitals. Our results also suggest that the availability of equally represented classes has enabled SVM to better characterise the data points.

Conclusion: The findings of the study presented here support the use of 3D TA on conventional MR images to aid diagnostic classification of paediatric brain tumours.

Keywords

3D texture analysis; radiomics; multicentre; MRI; pediatric brain tumors; classification; transferability; machine learning

List of Abbreviations (excluding standard)

AUC = area under the receiver operator characteristics curve
BCH = Birmingham Children's Hospital
CCLG = Children's Cancer and Leukaemia Group
DICOM= Digital imaging and communication in medicine
EP = ependymoma
GLCM = Grey-level co-occurrence matrix
GLRLM = Grey-level run-length matrix
GOSH = Great Ormond Street Hospital
LOOCV = Leave-one-out cross-validation
MDL = Minimum descriptive length
MB = medulloblastoma
NUH = Nottingham University Hospital
PA = pilocytic astrocytoma
ROC = receiver operator characteristics
ROI = Region of interest
SVM = support vector machine
SMOTE = synthetic minority over-sampling technique
TA = texture analysis

1. Background

Cancer is a leading cause of mortality from disease in children, with the latest available statistics in the UK showing that between 2009 and 2011, an average of 1,574 children per year were diagnosed with cancer, 16% of whom subsequently died [32]. Brain and central nervous system (CNS) tumours form the second most common group of cancers in children, accounting for 27% of all childhood cancers [32]. In order to tailor surgery and drug-based therapy, a brain tumour must be classified as one of a wide range of types, as per the scheme outlined by the World Health Organisation (WHO) [33].

Magnetic resonance imaging (MRI) is the key imaging technique used for visualising and managing paediatric brain tumours [1], [2], and initial assessment of tumours from MRI scans is usually performed qualitatively by radiologists [3]. However, the current gold standard for obtaining definite diagnosis is histopathological examination of biopsy samples taken through surgery [2],[4], because different brain tumour types do not always demonstrate clear differences in visual appearance [5], and using only conventional MRI to provide a diagnosis could potentially lead to inaccurate results [2].

The emerging field of *radiomics* provides a potential solution for non-invasive tumour characterisation by converting medical images into mineable data, through the extraction of a large number of quantitative imaging features [34], [35]. When developing quantitative medical image analysis techniques, it is usual to consider attributes which radiologists explicitly or implicitly use in their assessment of a specified tissue appearance. Intensity, morphology and texture are common examples of such important image attributes [36], [37]. Image texture can be defined as the spatial variation of pixel intensities within an image, and is known to be particularly sensitive for the assessment of pathology [36]. Visual assessment of texture is, however, particularly subjective. Additionally, it is known that human observers possess limited sensitivity to textural patterns, whereas computational texture analysis (TA) techniques can be significantly more sensitive to changes [36], [37].

A recent study reported by Fetit et al. [38] looked into the efficacy of MRI textural features, within a machine-learning framework, in diagnosing common paediatric brain tumour types. The study made use of datasets acquired from a single-centre and showed evidence that diagnostic classification can be optimised through the use of three-dimensional (3D) attributes, obtained through the analysis of multiple MR imaging slices. Despite the positive results reported in the paediatric and adult brain MRI literature, TA has not yet found its way into routine clinical practice. This is perhaps due to the sensitivity of textural features to variations in MR acquisition parameters, which may impede the transfer of results across various imaging centres [14]. In addition to this, the efficacy of TA is heavily dependent on the choice of textural features used to capture imaging patterns, which is linked to the choice of feature-selection methods used [14]. However, very little comparative work is available on studying the aforementioned issues [14].

Additionally, other factors continue to impede momentum. As argued by van Genneken et al. [42], for computer-aided-diagnosis systems to be useful for lesion detection and quantification applications, their stand-alone performance should be close to that of an expert radiologist. This is, however, rather challenging to evaluate when taking various aspects into account (e.g. economical, psychological and healthcare implications of additional examinations). Trying to prove a clinically relevant improvement adds additional complexity when the radiologist baseline performance is high (and sometimes varying) to start with. Also, there may be practical issues in terms of ensuring rapid TA so that results would be ready at the time of reporting for incorporation into radiological reports. Nevertheless, before such translation barriers can be addressed, one must show that TA is, in principle, a transferrable technique in paediatric oncology.

The study presented here expands the work discussed in [38] to include multicentric datasets obtained through a collaboration of three imaging centres across the UK. The primary aim of this study was to determine the efficacy and cross-centre transferability of 3D TA for non-invasive diagnostic classification of paediatric brain tumours from MR images. The study also aimed to investigate, through the use of supervised feature selection, the nature of features that are most likely to train classifiers that can generalise

well with the 3D textural data. Finally, the issue of class-imbalance, which arises due to some tumour types being more common than others, was investigated. To the best of the authors' knowledge at the time of writing, there are no published studies that used multicentric cohorts in order to assess the effectiveness and transferability of 3D MRI TA in paediatric oncology.

2. Materials and Methods

Figure 1 shows an overview of the experiment's set-up. Details are discussed below.

2.1 Patient Cohort

The clinical material used in this retrospective study consisted of pre-contrast T1 and T2-weighted MR images of 134 children with verified and untreated brain tumours. Forty-five were medulloblastomas (MB), seventy-one were pilocytic astrocytomas (PA) and eighteen were ependymomas (EP). In order to obtain diagnoses in accordance with the WHO classification, tumour samples were taken from all patients and underwent histopathological examination. Inclusion of datasets was not limited to tumours occurring in a certain location of the brain. Approval for the study was obtained from the research ethics committee, and informed consent was taken from parents or guardians.

2.2 Image Acquisition

Image acquisition was carried out in three UK centres: Birmingham Children's Hospital (BCH), Nottingham University Hospital (NUH) and Great Ormond Street Hospital (GOSH). The scanners used were 1.5T Siemens Symphony, 1.5T Siemens Avanto, 1.5T General Electric Signa, 1.5T Phillips Intera and 3T Phillips Achieva, and acquisition was carried out following a common protocol defined by the Children's Cancer and Leukaemia Group (CCLG) Functional Imaging Group. TE, TR, slice thickness and slice gap settings used in each imaging centre are summarised in Table 1. All images were anonymised and held at a secure e-repository provided by CCLG [15], [16].

2.3 Tumour Segmentation

In practice, tumour region of interest (ROI) outlining is usually performed manually, which is not only a time consuming task, but is also open to subjective interpretation of the radiologist. The medical imaging literature contains a plethora of work conducted on studying the effectiveness of automatic and semi-automatic segmentation algorithms. Whilst the implementation of an effective segmentation method is crucial for capturing tumour information, the details of which approach yields optimal segmentation is not of immediate interest within the context of this experiment.

Axial slices were manually chosen from each dataset using RadiAnt DICOM viewer (Medixant, Poland). Semi-automatic segmentation was performed on MATLAB (MathWorks, Massachusetts) using the snake gradient vector flow (Snake GVF) technique, as proposed by Xu and Prince [17], in order to extract the ROIs in which the tumour was present. The Snake GVF technique works by relying on manually defined seeding points, which are initially outlined by the user. The segmentation boundary is then constructed by calculating an edge map of the input image and progressing the contour towards a so-called force balance condition, where an internal force that prevents contour stretching is balanced with an external force that pulls the snake towards the desired contour.

The ROIs were checked visually to ensure that the segmentation technique worked sufficiently well. A single postgraduate research student in their second year of a doctorate programme at the time of carrying out the experiment performed the segmentation. Median segmented lesion sizes in cubic pixels were as follows, MB: 19,181 (3,123 – 36,775), PA: 24,313 (4,169 – 60,512), and EP: 19,488 (7,404 - 42,726).

2.4 ROI Normalisation

Although the data had been acquired under a standard acquisition protocol defined by the CCLG, there are usually variations in parameter settings that lead to the images having inconsistent grey-level ranges. Additionally, the three centres use scanning units that are obtained from different

manufacturers, and in the case of one centre, the scanners operate under different magnetic field strengths (1.5 T and 3 T). In order to mitigate the aforementioned issues, the segmented ROIs' grey-level intensities were normalised. Using MaZda texture analysis software (version 4.6) [8], normalisation was carried out through a two-step process that involved (1) grey-level range selection and (2) image quantisation. The first step was performed using *the limitation of dynamics to μ+/-3δ*, where μ is the mean grey-level value and δ is the standard deviation. This technique was shown to achieve reliable results on MR image texture classification by Collewet et al 2004 [18] and it was the same approach followed in the paediatric single centre study reported by Fetit et al 2015 [38]. Quantisation of the resulting grey-level range was done to compress it between 1 and 2^k, where k is the number of bits per pixel (k was chosen to be 6 bits for consistency with [38]).

2.5 Extraction of Textural Features

3D TA was carried out on the normalised T1 and T2-weighted ROIs using MaZda. We used multiple adjacent T1 and T2-weighted ROIs to calculate metrics that hold 'intra-slice' and 'inter-slice' pixel relationships. The TA methods used are based on histogram statistics, absolute gradient, grey-level co-occurrence matrix (GLCM) [19] and grey-level run-length matrix (GLRLM) [20].Multiple GLCMs and GLRLMs were computed along the 0°, 45°, 90°, 135° and z-axis directions; displacements chosen for computing GLCMs were 1, 2, 3 and 4 pixels (see Table 2 for summary). Mathematical definitions of the extracted features are included in the Supplementary material.

2.5 Feature Selection

The features extracted using MaZda were aggregated for further analysis, as per Table 3. The feature sets were imported to Orange (version 2.7), an open-source machine learning software library that was developed at the University of Ljubljana, Slovenia and allows data analysis through python scripting or intuitively via a graphical user interface [21].

As seen in Table 3, testing all possible combinations from all techniques, modalities and datasets would give a very large number of features (566). If all

of these features were evaluated together, it is very likely that our classification models will be over-fitted and poorly generalised [22], [23]. Irrelevant and redundant features are problematic because they may confuse the learning algorithm, by helping to obscure the distributions of the subset that holds influential features [23]. The number of features tested must therefore be reduced by *feature selection*. Feature selection also has the advantage of reducing training time and improving interoperability.

To this end, a number of feature selection algorithms were considered, the first being ReliefF [24]. The idea behind this technique is to estimate the effectiveness of a feature based on how well its value compares to that of the instance's neighbours. This is achieved by searching for an instance's nearest neighbours and finding an instance from the same class (nearest hit) and another one from a different class (nearest miss). The algorithm then uses a weighting approach to estimate the quality for each feature. Good features are assumed to have the same value for instances from the same class and should discriminate between instances that belong to different classes [24].

The entropy minimum descriptive length (MDL) discretisation technique, which is usually used to partition continuous features to a discrete number of intervals, was also considered for feature selection. Since a feature's entropy can be used as a measure of its discriminative power, entropy-based discretisation can also be used for feature selection [25]. Our discretised feature sub-set therefore holds only the features that the algorithm deduced to be the most relevant and discriminative.

Thirdly, we were also interested in studying the use of features selected using a feature selection pipeline, comprising a combination of both algorithms: Entropy-MDL and ReliefF.

2.6 Classification Model

Using python's Orange library, we designed a cost-based support vector machine (C-SVM) classifier that used the radial basis function (RBF) kernel and a cost coefficient of 1, to be trained with textural features. A detailed review discussing the principles behind the SVM classification algorithm can be found at [26].

It is worth noting that the single-centre study conducted by Fetit et al [38] investigated 6 different models that represent typical implementations of popular classification algorithms, and suggested that the choice of classifier is not of substantial importance for this particular problem of paediatric brain tumour diagnosis. Thus, comparative analysis of different classification algorithms was not of immediate interest within this multicentric work, and we chose to focus on analysing the behaviour of a typical SVM implementation.

2.7 Model Validation

(a) Examining Transferability by Pairwise Testing by Hospital on Unseen Data

To determine the practical influence of differences in textural feature-sets extracted from different MRI centres, three different instances of the SVM classifier were created, each being trained on features extracted from one of the three hospitals. Assessing how each SVM performed on unseen datasets, which were obtained from the other two hospitals, was carried out for testing. For example, the performance of an SVM trained with Birmingham Children's Hospital data was evaluated by testing on datasets obtained from Great Ormond Street Hospital and Nottingham University Hospital.

The metric used for primary evaluation of performance was the area under the receiver operating characteristics (ROC) curve, or simply AUC. The predictive ability of a classifier is typically measured by its predictive accuracy or by its error rate, which is one minus the accuracy. However, recent research efforts in machine learning have shown that AUC is a more consistent and discriminant measure of classification performance than accuracy [27], [28]. Advantages of using AUC as a primary measure of performance include its insensitivity to imbalanced class distributions [29]. This is relevant to our

dataset because of the lack of a sufficient number of Ependymoma samples, which form only 13% of the multicentre cohort. In addition to this, it has been shown that if a classifier is optimised to yield maximum AUC, it is also likely to perform well in the accuracy measure, whereas classifiers that are optimised to yield high accuracy do not necessarily perform well in the AUC measure [27] [28]. Nevertheless, we additionally reported on the observed classification accuracy obtained in order to ease comparison with the single-centre work previously reported by Fetit et al in 2015 [38].

(b) Estimating Overall Performance Using Leave-One-Out Cross-Validation

Testing the models on unseen data mainly aimed to examine the cross-centre transferability of TA. In order to get an overall estimate of the models' classification performance, Leave-One-Out Cross-Validation (LOOCV) was additionally carried out on an aggregated feature-set comprising data from all three hospitals (all 134 samples). AUC, classification accuracy, sensitivity and specificity were measured from the results. 95% confidence intervals of overall classification accuracies were calculated using bootstrapping (1000 samples were generated). Since the aim of this step was not to investigate optimal settings for classification, but to get an estimate of the overall performance, only one feature selection method (Entropy-MDL) was used.

(c) Addressing the Class Imbalance Problem

A dataset is considered imbalanced if the classes are not approximately equally represented. The machine learning community has tackled the class imbalance issue in two ways. One is to assign costs to training examples, and the other is by over sampling minorities (or under sampling majorities). Although the data used for this study had been acquired at three different hospitals, the classes represented are quite imbalanced in the sense that EP forms only 13% of the overall dataset (18/134). This may be problematic because the minority samples might be ignored by the classifier, and potentially leading to poor EP sensitivity.

In order to investigate this, a separate analysis was carried out where the synthetic minority over-sampling technique (SMOTE) was applied to the

extracted 3D features. SMOTE was used to create 27 synthetic EP samples by operating in feature space. This method works by taking each minority class sample and introducing synthetic examples along the line segments joining any/ all of the k minority class nearest neighbours. The neighbouring points are randomly chosen depending on the amount of over-sampling required. Using an SVM classifier, LOOCV was carried out on the new feature-set comprising 161 samples. It is important to note that all SMOTE cases were included in the LOOCV loop. Classification accuracy, sensitivity, specificity and AUC were measured from the test results. 95% confidence intervals of classification accuracies were calculated using a bootstrapping of subjects in the sampling (1000 samples were generated).

3 Results

(a) Results from Pairwise Testing by Hospital on Unseen Data

Table 4 and Figure 2 show a summary of AUC values obtained with SVM classifier. The mean AUC values obtained by Entropy-MDL, ReliefF and the hybrid pipeline are 74.5%, 71.8% and 76% respectively. The highest AUC value was obtained when SVM was trained on NUH data and tested on BCH data (86% on ReliefF) - an interesting finding since one of the scanners in NUH uses magnetic field strength of 3T, whereas both scanners used to acquire BCH data are 1.5T.

For each feature selection method, the number of chosen features that were required to yield optimum AUC are reported in Table 5. It is important to note that for pairwise testing, feature selection was separately carried out on the corresponding training set for each of the 6 tests, and not on the entire dataset. Whilst with entropy-MDL no manual definition of feature percentages is performed for the algorithm to operate, we reported the number of features that were discretised and hence deemed important by the algorithm.

The optimal features identified for each of the six tests are reported in Tables 6 and 7. It is worth noting that there was a common set of attributes that were deemed as optimal in all six tests (e.g. sum of squares, sum average and difference entropy). However, the particular feature variation (i.e. the specific pixel distance and direction) that was identified as important varied greatly across the six tests. *Please refer to the Supplementary Material for an explanation of the features and their offsets*. For instance, even though the sum of squares attribute was identified as an important feature in both tests 1 and 2, test 1 showed that offsets (0,0,3) and (0,0,4) were important, whereas test 2 showed that offsets (1,0,0), (0,1,0) and (0,0,2) were important. Besides reporting on AUC, and to facilitate comparison with the single-centre study reported previously by Fetit et al [38], we also reported on the classification accuracies yielded when entropy-MDL was used. Those ranged between 47% and 64%.

(b) LOOCV results on the original dataset (before over-sampling)

Table 8 lists the results obtained when the entire feature-set, comprising all 134 samples, was tested with an SVM classifier using LOOCV. Results were generally satisfactory, with the overall AUC being 86% (see Figure 3 for ROC curves). The corresponding overall classification accuracy was also computed, to ease comparison to the current state-of-the art in the literature, and was found to be 72%. However, it is worth noting that EP demonstrated a very low sensitivity value of 11%.

(c) LOOCV results after minority over-sampling

Table 9 lists the LOOCV results obtained when an SVM classifier was used on the new feature-set that comprised an additional 27 (synthetic) EP samples (see Figure 4 for ROC curves). The noticeable increase in EP sensitivity (from 11% to 87%) suggests that the availability of equally represented classes has enabled SVM to better characterise the data points. It is important to note that the SMOTE cases were included in the LOOCV loops.

3. Discussion

This work presented a multicentre investigation on the efficacy and transferability of using multi-slice (3D) statistical textural features extracted from conventional MR images, within a machine-learning framework, to discriminate between the most frequently occurring paediatric brain tumours: medulloblastoma, pilocytic astrocytoma and ependymoma. The study made use of standard pre-contrast T1 and T2-weighted images, which are routinely acquired when children present with suspected brain tumours. For the purpose of this discussion, the main areas of interest were:

1. Do the classification results show enough evidence that 3D TA is a transferrable technique, allowing for its use across multiple centres?
2. If so, are there conditions that need to be taken into account prior to clinical translation of 3D TA?
3. What is the nature of features deduced to be optimal as per feature selection?

With regards to the individual pair-wise tests, the primary metric that was evaluated when looking for optimal classifier performance was AUC. Note that the statistical meaning of AUC can be defined as the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. For each of the tests reported in Table 4, optimal AUC ranged between 76% and 86%, which suggests that the use of three-dimensional textural features generally enabled SVM to capture transferable tumour information that could be used to successfully classify images obtained from other imaging centres.

Classification accuracies obtained with entropy-MDL, however, ranged between 47% and 64%. It can be deduced from the obtained classification accuracies and optimal AUC values, that the performance is considerably weaker than what was previously reported in the single-centre study on paediatric 3D TA [38], which in some settings reached over 95% for both metrics. This suggests that for successful long-term clinical translation of TA, classifiers will be more likely to make effective predictions when trained and applied on datasets acquired using the same scanners.

It is also interesting to note that the performance obtained when using the entire dataset (both before and after SMOTE) was generally more effective than the findings of the individual pair-wise tests; which further emphasises the need for training and test-sets to have data acquired using the same scanners if TA is to be used in practical settings.

Feature selection results suggest that similar aspects of tumour texture are enhanced by MR images obtained at different hospitals, since a common set of attributes was identified as important in all six pairwise-testing tests. Such attributes include *sum of squares*, *sum average* and *difference entropy*. However, the particular variation of distance and direction of analysis varied across the six tests and was heavily reliant on the test-sets used, even when features were extracted from the same centre.

The above observations suggest that whilst TA is, in principle, a scalable technique that can be used to classify tumour types across different hospitals,

there does not seem to be enough indication of any 'universal' features that could be measured across specific directions and distances of analysis for use across centres, without taking other factors into account.

The dependency of optimal performance on both acquisition scanners used and test-sets also suggests that for TA to be used in practice, there needs to be a robust means of selecting representative features for training, which is likely to vary depending on each individual scenario. One potential solution is to combine the attributes that were identified as important across tests into a single score, perhaps through averaging, as a means of decreasing any inherent noise, increasing robustness and improving reliability. Additionally, meta-analysis of the performance of different feature selection methods needs to drive future efforts in this area.

Besides the single-centre work reported by Fetit el at [38], the closest work to this experiment is the multicentre study by Tantisatirapong et al [39], where conventional 2D TA was used in a binary classification problem to diagnose paediatric MB and PA, yielding an overall classification accuracy of 77% using T2-weighted data. Whilst it is not possible to directly compare the findings of this experiment to the current state-of-the-art, due to variations in primary aims and methodologies, the overall classification accuracies of 72% (before SMOTE) and 77% (after SMOTE) are in line with what had been reported in [39].

Although 3D TA of MRI relies heavily on sophisticated mathematical procedures, this study was entirely carried out using commercially available and open-source software, which are provided with well-documented manuals to support their use by personnel with limited programming backgrounds.

4. Study Limitations and Future Work

This study suffers from the limitation that the presented pairwise testing

results are a best-case scenario, as the comparison looked into optimal AUC values. In terms of feature selection, a limitation to the LOOCV work was that the feature selection algorithm was applied prior to the validation loop, and not within each validation fold, which might have introduced an element of bias to our overall estimate of classifier performance. Another limitation is that synthesised samples generated by SMOTE were included in the testing folds, and ideally the assessment of performance needs to be only focused on the original samples.

In order to assess robustness of the obtained findings, it will be necessary to further test the classifiers with the optimal settings identified in this study. This can be done using a three-fold validation approach, where training is done on one dataset, followed by a testing stage on another dataset where the optimal classifier settings are identified, and finally a validation stage where the identified optimal settings are tested for robustness. Although the analysis was carried out on the three most frequently occurring paediatric brain tumours (MB, PA and EP), this methodology can be extended to other brain tumour types, provided enough data samples are available for use as a test-set. It will also be interesting to look into the use of 3D TA on diffusion- weighted imaging (DWI), as work currently available in the literature has shown promising results with conventional 2D TA of DWI.

It is also worth noting that, as can be seen in Table 1, a fundamental limitation of this work was the clear heterogeneity in acquisition settings of the MR scans, which were used retrospectively. Such intrinsic variations in TE, TR and slice separation can introduce high levels of noise in the information captured by the textural features; and it is likely that such noise could have contributed to the lack of a 'universal' offset in important textural features.

3D TA would ideally require minimal MR image slice gaps to maximise information captured by volumetric features. This is challenged by the clinical datasets used here, which were retrospective multi-slice MR scans acquired using conventional Spin-Echo sequences (i.e. the images are not true 3D). Additionally, the use of slices for image acquisition means that each slice summarises potentially many different elements of the underlying pathological

structure over its width [41]. This leads to the interesting question of whether selecting thinner slices during acquisition might result in images that can build more robust TA predictive models (the use of thinner slices would, however, result in a worsened signal-to-noise ratio, thus concealing the true texture). In this regard, it is likely that between-plane textural features that were calculated via 3D TA are lacking in robustness.

In the authors' opinion, a major limitation that MRI texture analysis studies suffer from – one that impedes clinical translation - is the lack of clear clinical meanings to the imaging features identified as biomarkers. Establishing such meaning is a challenging task since TA, in theory, captures underlying MR imaging patterns that are below human vision. One way this issue could be investigated in future work is by carrying out TA on histopathological samples under different microscopic scales, where clinical attributes can be easily correlated with important textural features. Assuming that such meaning could be translated to MR imaging scales, this could potentially provide radiologists with a number of textural patterns to look for when carrying out initial tumour characterisation. Good understanding of feature meanings will ensure that the generated knowledge and the explanation of classifier decisions will be transparent to the clinicians. This will support clinical acceptance of TA, since according to Kononenko [40], transparency is an important requirement for decision-support systems to be useful in solving medical diagnostic tasks.

Among the reasons for slow acceptance of decision support systems in clinical settings, perhaps the most reasonable one is that the introduction of such technologies will further increase the abundance of tools and instrumentation available to clinicians [40]. The use of non-invasive TA would have the undesirable side effect of further increasing the complexity of the radiologist's work, which is already sufficiently complicated. Therefore, TA and machine learning systems will have to be integrated into the existing instrumentation that makes its adoption as natural as possible.

5. Conclusions

The results of the study presented here indicated that despite the differences in textural information among MR images from different hospitals, feature-sets

from one hospital may be used for successful tumour type classification when tested on data from other hospitals; an important finding for future clinical adoption of TA. The findings of the study presented here support the use of 3D TA on conventional MR images to aid diagnostic classification of paediatric brain tumours.

Acknowledgements

AEF and TNA would like to thank Warwick Manufacturing Group (WMG) at the University of Warwick for financial support. JN is supported by Help Harry Help Others. ACP is supported by a National Institute of Health Research (NIHR) Research Professorship (13–0053). The authors would like to thank Ramneek Kaur for maintaining entries to the CCLG database. All authors acknowledge the support received from Cancer Research UK (CRUK) and the Engineering and Physical Sciences Research Council (EPSRC) Cancer Imaging Programme at the Children's Cancer and Leukaemia Group (CCLG) in association with the Medical Research Council (MRC) and Department of Health (England) (C7809/A10342).

References

[1] A. C. Peet, T. N. Arvanitis, D. P. Auer, et al, "The value of magnetic resonance spectroscopy in tumour imaging.," *Arch. Dis. Child.*, vol. 93, no. 9, pp. 725–727, Sep. 2008.

[2] J. Vicente, E. Fuster-Garcia, S. Tortajada, et al "Accurate classification of childhood brain tumours by in vivo ^{1}H MRS - a multi-centre study.," *Eur. J. Cancer*, vol. 49, no. 3, pp. 658–67, Feb. 2013.

[3] A. Kassner and R. E. Thornhill, "Texture Analysis : A Review of Neurologic MR Imaging Applications," *AJNR. Am. J. Neuroradiol.*, no. May, pp. 809–816, 2010.

[4] L. M. Harris, N. Davies, L. Macpherson, et al, "The use of short-echo-time 1H MRS for childhood cerebellar tumours prior to histopathological diagnosis.," *Pediatr. Radiol.*, vol. 37, no. 11, pp. 1101–9, Nov. 2007.

[5] E. Orphanidou-Vlachou, N. Vlachos, N. P. Davies, T. N. Arvanitis, R. G. Grundy, and A. C. Peet, "Texture analysis of T1 - and T2 -weighted MR images and use of probabilistic neural network to discriminate posterior fossa tumours in children.," *NMR Biomed.*, Apr. 2014.

[6] M. D. C. Alegro, S. Y. Bando, A. V. Silva, and B. C. Medeiros, "Texture Analysis and Classifi ers Applied to High-Resolution MRI from Human Surgical Samples in Refractory Mesial Temporal Lobe Epilepsy," pp. 40–43, 2012.

[7] D. Rodriguez Gutierrez, A. Awwad, L. Meijer, et al, "Metrics and Textural Features of MRI Diffusion to Improve Classification of Pediatric Posterior Fossa Tumors.," *AJNR. Am. J. Neuroradiol.*, p. ajnr.A3784–, Dec. 2013.

[8] A. M. and A. K. M. Strzelecki, P. Szczypinski, "A software tool for automatic classification and segmentation of 2D/3D medical images," *Nucl. Instruments Methods Phys. Res.*, vol. 702, pp. 137–140, 2013.

[9] P. Georgiadis, D. Cavouras, I. Kalatzis, et al, "Enhancing the discrimination accuracy between metastases, gliomas and meningiomas on brain MRI by volumetric textural features and ensemble pattern recognition methods.," *Magn. Reson. Imaging*, vol. 27, no. 1, pp. 120–130, Jan. 2009.

[10] D. Mahmoud-Ghoneim, G. Toussaint, J. M. Constans, and J. D. de Certaines, "Three dimensional texture analysis in MRI: a preliminary evaluation in gliomas.," *Magn. Reson. Imaging*, vol. 21, no. 9, pp. 983–987, Nov. 2003.

[11] W. Chen, M. L. Giger, H. Li, U. Bick, and G. M. Newstead, "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images.," *Magn. Reson. Med.*, vol. 58, no. 3, pp. 562–571, Sep. 2007.

[12] P. Georgiadis, S. Kostopoulos, D. Cavouras, et al, "Quantitative combination of volumetric MR imaging and MR spectroscopy data for the discrimination of meningiomas from metastatic brain tumors by means of pattern recognition.," *Magn. Reson. Imaging*, vol. 29, no. 4, pp. 525–35, May 2011.

[13] A. E. Fetit, J. Novak, A. C. Peet, and T. N. Arvanitis, "3D texture analysis of MR images to improve classification of paediatric brain tumours: a preliminary study.," *Stud. Health Technol. Inform.*, vol. 202, pp. 213–6, Jan. 2014.

[14] M. E. Mayerhoefer, M. J. Breitenseher, J. Kramer, N. Aigner, S. Hofmann, and A. Materka, "Texture analysis for tissue discrimination on T1-weighted MR images of the knee joint in a multicenter study:

Transferability of texture features and comparison of feature selection methods and classifiers.," *J. Magn. Reson. Imaging*, vol. 22, no. 5, pp. 674–80, Nov. 2005.

[15] J. Rossiter, T. Arvanitis, K. Natarajan, et al, "A clinical trials e-repository with integrated conventional and functional imaging data," Jun. 2012.

[16] T. Arvanitis, K. Natarajan, J. Rossiter, et al, "THE CHILDREN'S CANCER AND LEUKAEMIA GROUP (CCLG) FUNCTIONAL IMAGING E-REPOSITORY FOR CLINICAL TRIALS OF CHILDHOOD BRAIN TUMOURS," Jun. 2010.

[17] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow.," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 359–69, Jan. 1998.

[18] G. Collewet, M. Strzelecki, and F. Mariette, "Influence of MRI acquisition protocols and image intensity normalization methods on texture classification.," *Magn. Reson. Imaging*, vol. 22, no. 1, pp. 81–91, Jan. 2004.

[19] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Syst. Man. Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[20] M. M. Galloway, "Texture analysis using gray level run lengths," *Comput. Graph. Image Process.*, vol. 4, no. 2, pp. 172–179, 1975.

[21] J. Demšar, T. Curk, A. Erjavec, et al, "Orange: data mining toolbox in python," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 2349–2353, Jan. 2013.

[22] M. Dash and H. Liu, "Feature Selection for Classification," *Intell. Data Anal. 1*, no. 97, pp. 131–156, 1997.

[23] G. Koller, "Toward Optimal Feature Selection," *Tech. Report. Stanford Infolab.*

[24] I. Kononenko and E. Simec, "Induction of decision trees using ReliefF," *Proc. ISSEK94*, no. 363, pp. 199–220.

[25] J. Liu and L. Wong, "Mean-entropy Discretized Features are Effective for Classifying High-dimensional Bio-medical Data," *Citeseer*.

[26] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: a review of classification and combining techniques," *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, Nov. 2007.

[27] Charles X. Ling, J. Huang, and H. Zhang, "AUC: a statistically consistent and more discriminating measure than accuracy," *Proc. 18TH Int. Conf. Artif. Intell.*, 2003.

[28] C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, Mar. 2005.

[29] T. Fawcett, "ROC Graphs: Notes and Practical Considerations for Researchers," *ReCALL 31*, 2004.

[30] G. Castellano, L. Bonilha, L. M. Li, and F. Cendes, "Texture analysis of medical images.," *Clin. Radiol.*, vol. 59, no. 12, pp. 1061–9, Dec. 2004.

[31] W. H. Nailon, "Texture Analysis Methods for Medical Image Characterisation," *Biomed. Imaging Tech.*, pp. 76–100, 2004.

[32] Cancer Research UK. Childhood Cancer Statistics [Online].
Available http://www.cancerresearchuk.org/cancer-info/cancerstats/childhoodcancer/

[33] Louis, D.N., Perry, A., Reifenberger, G. et al, "The 2016 World Health Organization Classification of Tumours of the Central Nervous System: a summary", Acta Neuropathol (2016) 131: 803. doi:10.1007/s00401-016-1545-1

[34] V. Kumar, Y. Gu, S. Basu, A. Berglund, et al, "Radiomics: the process and the challenges," Magnetic Resonance Imaging, vol. 30, no. 9, pp. 1234-48, 2013.

[35] C. Parmar, E. R. Velazquez, R. Leijenaar, et al, "Robust Radiomics Feature Quantification Using Semiautomatic Volumetric Segmentation," PLoS One, 9:e102107, 2014.

[36] R. A. Lerski, K. Straughan, L. R. Schad, D. Boyce, S. Blml, and I. Zuna, "MR image texture analysis - An approach to tissue characterization," Magnetic Resonance Imaging, vol. 11, no. 6, pp. 873-887, Jan. 1993.

[37] A. Bruno, R. Collorec, J. Bezy-Wendling, P. Reuze, and Y. Rolland, "Texture Analysis in Medical Imaging," Studies in Health Technology and Informatics vol. 30, pp. 133-164, 1997.

[38] A. E. Fetit, J. Novak, A. C. Peet and T. N. Arvanitis, "Three-dimensional textural features of conventional MRI improve diagnostic classification of childhood brain tumours", NMR in Biomedicine, vol. 28, no. 9, 1174-1184, 2015.

[39] S. Tantisatirapong, N. P. Davies, L. Abernethy, et al, "Automated Processing Pipeline for Texture Analysis of Childhood Brain Tumours based on Multimodal Magnetic Resonance Imaging," Biomedical Engineering, vol. 791, pp.791-081, 2013.

[40] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective," Artificial Intelligence in Medicine, vol. 23, pp. 89-109, 2001.

[41] M. Hajek, “Texture Analysis for Magnetic Resonance Imaging,” European Network Cost Action B21, 2006.

[42] B. van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic", Radiology, vol. 261 (3), 719-32, 2011.

TABLES

	BCH	NUH	GOSH
Scanners	GE Signa 1.5 Tesla and Siemens Symphony 1.5 Tesla	Phillips Achieva 3.0 Tesla and Phillips Intera 1.5 Tesla	Siemens Avanto 1.5 Tesla and Siemens Symphony 1.5 Tesla
T1 TE	8.4-22 ms	12-23 ms	2.39 – 51 ms
T1 TR	360-819 ms	373 -1400 ms	212.57 – 697 ms
T1 Slice Thickness	4-5 mm	4-5 mm	4.5-5 mm
T1 Slice Gap	0.8-1.5 mm	4.4-6.5 mm	1-3 mm
T2 TE	77-105 ms	83.5-125 ms	14-115 ms
T2 TR	3940-7840 ms	3000-6475 ms	3530 – 6180 ms
T2 Slice Thickness	3-5 mm	4-5 mm	3.5 -5 mm
T2 Slice Gap	0.6-1.5 mm	0.4-1.5 mm	1-3.5 mm

Table 1 Scanner models, field strengths and acquisition settings for the three imaging centres.

TA method	*Calculated features*
Histogram statistics	Mean, variance, skewness, kurtosis, minimum, maximum and percentiles (1%, 10%, 50%, 90% and 99%).
Absolute gradient statistics	Absolute gradient mean, variance, skewness and kurtosis
Grey-level co-occurrence matrix (GLCM)	Ang. Second Moment, inverse difference moment, contrast, correlation, entropy, sum of squares (variance), sum average, sum variance, difference variance and difference entropy
Grey-level run-length matrix (GLRLM)	Short run emphasis, long run emphasis, grey-level non-uniformity, run length non-uniformity and run percentages

Table 2 Summary of the TA methods used and their corresponding features

TA method	*T1*	*T2*
GLCM	240	240
GLRLM	24	24
Histogram	13	13
Absolute Gradient	6	6

Table 3 Summary of the number of features for each dataset

Test ID	*Training set*	*Testing set*	*Entropy-MDL optimal AUC*	*ReliefF optimal AUC*	*Pipeline optimal AUC*	*Optimal feature-selection Method*	*CA with entropy-MDL*
1	BCH	NUH	74%	83%	73%	ReliefF	64%
2	BCH	GOSH	74%	62%	80%	Pipeline	53%
3	NUH	BCH	74%	86%	74%	ReliefF	53%
4	NUH	GOSH	71%	85%	75%	ReliefF	63%
5	GOSH	BCH	76%	60%	76%	Entropy-MDL / Pipeline	47%
6	GOSH	NUH	78%	55%	78%	Entropy-MDL / Pipeline	64%

Mean AUC value:

Entropy-MDL: **74.5%**

ReliefF: **71.8%**

Pipeline: **76%**

Table 4 AUC values, in %, obtained through multicentre classification. Additionally, the classification accuracy (CA) obtained with entropy-MDL was reported on, to ease comparisons with the single-centre study previously reported in the literature. It is interesting to note that NUH datasets included images acquired using both 3T and 1.5T scanners, whereas the other two centres only included 1.5T scans. The performance of the pair-wise tests, therefore, suggest potential transferability of textural features across scanners of different strengths.

Test ID	*Training Set*	*Testing Set*	*ReliefF*	*Pipeline*
1	BCH	NUH	44	43
2	BCH	GOSH	50	19
3	NUH	BCH	33	28
4	NUH	GOSH	34	30
5	GOSH	BCH	100	14
6	GOSH	NUH	100	14

Table 5 Number of top-ranked features that were needed to yield optimal AUC performance.

Feature name	*Offset*	
Test 1	**T1**	**T2**
Angular Second Moment	0,1,0; 0,2,0; 0,0,2; 0,3,0	0,4,0
Contrast	0,0,4	
Difference Entropy	0,0,3; 0,0,4	0,2,0; 2-2,0; 3,-3,0; 0,3,0; 0,4,0; 4,-4,0
Difference Variance	0,0,4	0,1,0; 0,4,0
Entropy	0,0,3; 0,0,4	0,0,1
Fraction		45, 135 degrees; Vertical
Histogram	Skewness	Skewness
Inverse Difference Moment	0,1,0; 0,2,0; 0,3,0; 0,4,0	0,2,0; 0,3,0; 0,4,0; 4,-4,0
Long Run Emphasis	Vertical	
Short Run Emphasis		135 degrees; vertical; horizontal
Sum Average	0,0,3; 0,0,4	0,1,0
Sum Entropy	0,0,3; 0,0,4	0,0,1
Sum of Squares	0,0,3; 0,0,4	
Test 2	**T1**	**T2**
Angular Second Moment	0,0,3	
Difference Entropy	0,3,0	
Entropy	0,0,3	
Gradient		Non Zeros
Histogram		Max; Min; 50%; 90%; 99%; Mean; Variance; Kurtosis
Inverse Difference Moment	3,-3,0	4,-4,0
Sum Average		1,0,0; 0,02
Sum of Squares	1,0,0; 0,1,0; 0,0,2	
Test 3	**T1**	**T2**
Angular Second Moment	0,0,4	0,0,4
Contrast	0,0,4	0,0,4
Correlation	0,0,4	0,0,4
Sum of Squares	0,0,4	1,0,0; 1,1,0; 0,1,0; 1,-1,0; 0,2,0; 0,0,4
Inverse Difference Moment	0,0,4	0,0,4
Sum Average	0,0,4	0,3,0; 0,1,0; 0,2,0; 0,0,4
Sum Variance	0,0,4	0,0,4
Sum Entropy	0,0,4	0,0,4
Entropy	0,0,4	0,0,4
Difference Variance	0,0,4	0,0,4
Difference Entropy	0,0,4	0,0,4
Volume	0,0,4	0,0,4
Histogram	0,0,4	Skewness

Table 6 A table listing the optimal textural features identified in Tests 1 to 3.

Feature name	*Offset*	
Test 4	**T1**	**T2**
Angular Second Moment	0,0,4	0,0,4
Contrast	0,0,4	0,0,4
Correlation	0,0,4	0,0,4
Sum of Squares	0,0,4	0,1,0; 1,0,0; 1,1,0; 1,-1,0; 0,2,0; 0,0,4
Inverse Difference Moment	0,0,4	0,0,4
Sum Average	0,0,4	0,1,0; 0,2,0; 0,3,0; 0,4,0; 0,0,4
Sum Variance	0,0,4	0,0,4
Sum Entropy	0,0,4	0,0,4
Entropy	0,0,4	0,0,4
Difference Variance		0,0,4
Difference Entropy	0,0,4	0,0,4
Volume	0,0,4	0,0,4
Histogram		Skewness
Test 5	**T1**	**T2**
Correlation		0,0,1
Sum of Squares		0,0,2
Inverse Difference Moment		2,2,0; 2,-2,0; 4,-4,0
Sum Average		0,0,2; 0,1,0; 0,2,0
Sum Variance		0,1,0
Difference Entropy	0,0,1	3,-3,0; 4,-4,0
Histogram		Kurtosis
Test 6	**T1**	**T2**
Correlation		0,0,1
Difference Entropy	0,0,1	3,-3,0; 4,-4,0
Histogram		Kurtosis
Inverse Difference Moment		2,2,0; 2,-2,0; 4,-4,0
Sum Average		0,1,0; 0,2,0; 0,0,2
Sum of Squares		0,0,2
Sum Variance		0,1,0

Table 7 A table listing the optimal textural features identified in Tests 4 to 6.

	MB		*PA*		*EP*				
Overall AUC	Sens	Spec	Sens	Spec	Sens	Spec	Overall Accuracy	Variance of Accuracy	95% CI of Accuracy
86%	67%	82%	90%	71%	11%	97%	72%	0.0015	50%-84%

Table 8 A table listing the results obtained when the feature-set, comprising data from all three hospitals (134 samples), was tested with an SVM classifier on LOOCV. Entropy-MDL was used for feature selection. Accuracy, sensitivity and specificty are referred to as Acc, Sens and Spec respectively. Variance of over-all accuracy was calculated, with the assumption of a Binomial approximation to the count of correct classification, as p(1-p)/N, where p is the probability of correct classification and N is the number of samples. 95% confidence intervals for the overall classification accuracies were obtained by bootstrapping.

	MB		*PA*		*EP*				
Overall AUC	Sens	Spec	Sens	Spec	Sens	Spec	Overall Accuracy	Variance of Accuracy	95% CI of Accuracy
92%	57%	91%	83%	83%	87%	91%	77%	0.0011	60%-90%

Table 9 A table listing the classification results obtained with LOOCV, after SMOTE was applied to generate 27 synthetic EP samples. Entropy-MDL was used for feature selection. Accuracy, sensitivity and specificty are referred to as Acc, Sens and Spec respectively. Variance of over-all accuracy was calculated, with the assumption of a Binomial approximation to the count of correct classification, as p(1-p)/N, where p is the probability of correct classification and N is the number of samples. 95% confidence intervals for the overall classification accuracies were obtained by bootstrapping.

List of Figure Legends

Figure 1: Flowchart showing methodological overview of our experimental set up. It is important to note that for the pair-wise testing part of the experiment, the training and testing steps were part of an optimisation process, where the main interest was to examine what the ultimate optimal performance that could be achieved with the available cohort is. Hence, this study suffers from the limitation that the presented pairwise testing results are a best-case scenario. Ideally, three-fold validation should be used, where optimisation is carried out on two folds, and testing of performance is carried out on the third one. This is, however, not practical to implement here due to data size limitations.

Figure 2: Bar chart showing optimal AUC values obtained through pairwise testing for multicentre classification.

Figure 3: ROC curves depicting SVM classifier performance using the LOOCV scheme. All 134 samples obtained from three hospitals were used for the analysis. (a)Medulloblastoma (b)Pilocytic Astrocytoma and (c) Ependymoma. Overall AUC value is 86%.

Figure 4: ROC curves depicting SVM classifier performance using the LOOCV scheme, after SMOTE was used to generate 27 synthetic ependymoma samples. (a)Medulloblastoma (b)Pilocytic Astrocytoma and (c) Ependymoma. Overall AUC value is 92%